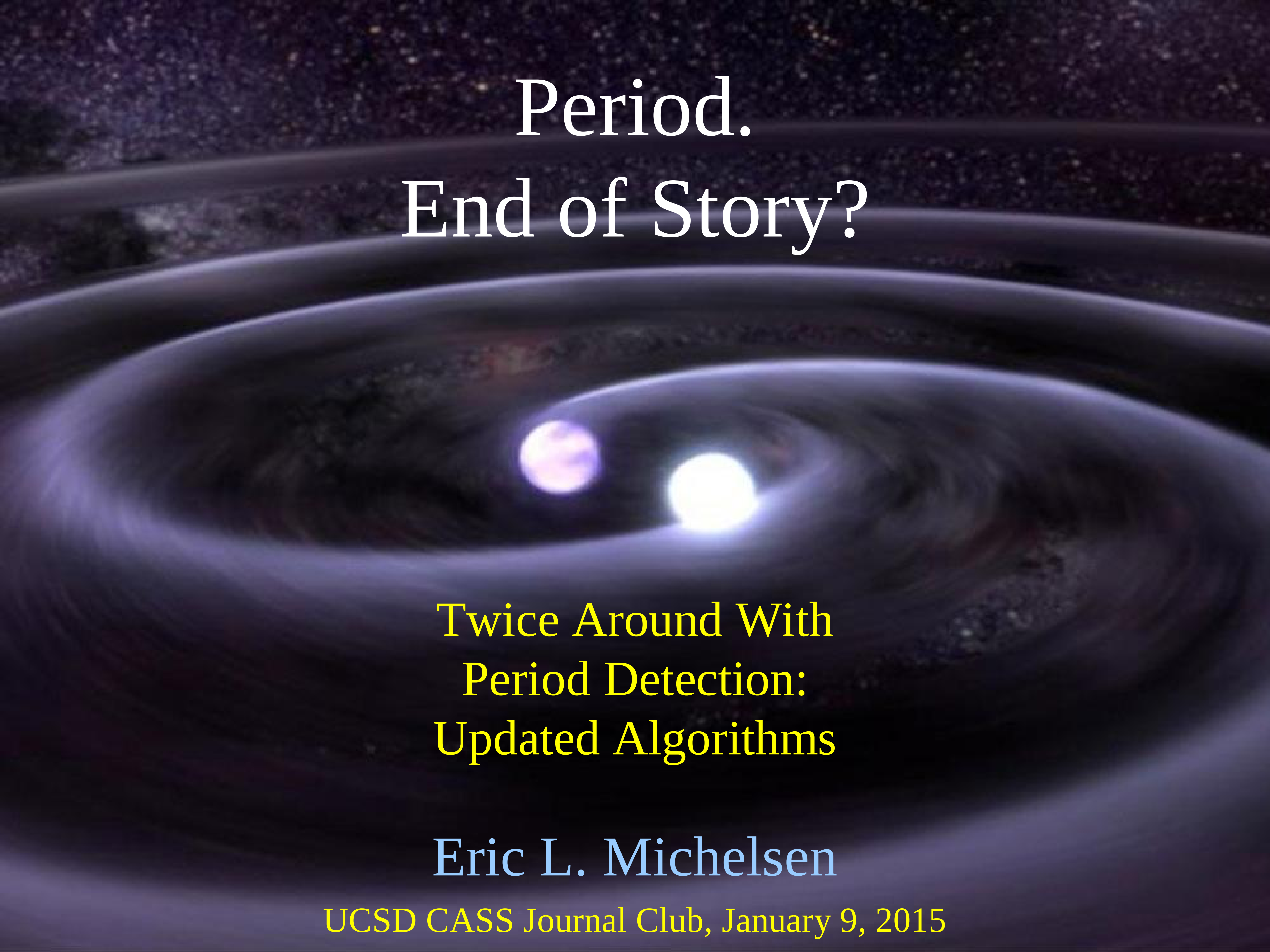


Period.
End of Story?

Twice Around With Period Detection: Updated Algorithms

Eric L. Michelsen

UCSD CASS Journal Club, January 9, 2015

Can you read this?

- 36 pt This is big stuff
- 24 pt This makes a point
- 20 pt This is very legible
- 18 pt This may be the minimum usable
- 16 pt For reference only
- 12 pt Are you kidding me?
- 10 pt Don’t even try it

Papers

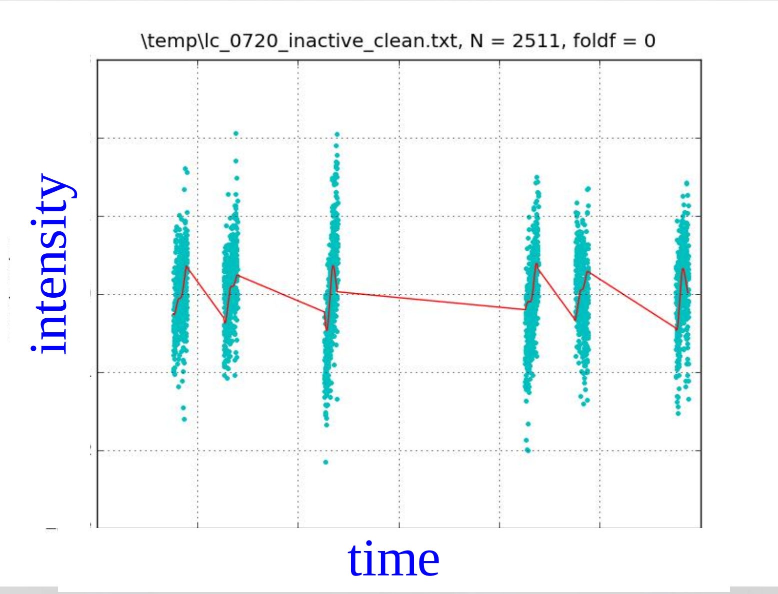
- Zechmeister, M. and M. K¨urster, “The generalised Lomb-Scargle periodogram”, Astronomy & Astrophysics, January 20, 2009
- Swarzenberg-Czerny, “The distribution of empirical periodograms: Lomb–Scargle and PDM spectra”, A. Schwarzenberg-Czerny”, Monthly Notices of the Royal Astronomical Society, 301, 831–840 (1998), and several similar.
- Some of my own work:
  - http://physics.ucsd.edu/~emichels/, *Funky Mathematical Physics Concepts*, 2014.

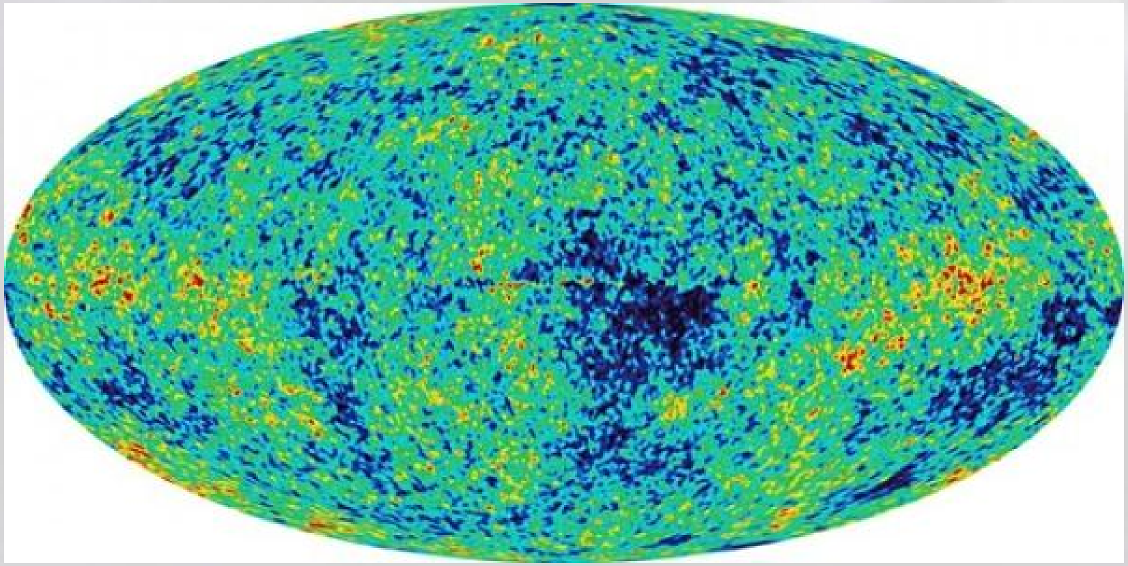
My one sentence

- Detecting periods is harder than you think
  - 70's and 80's algorithms are out of date
    - L-S, PDM, EF, (DFT was already unusable)
    - Seriously suboptimal
    - Standard formulas incorrect
  - Updates exist: join the 21st century
    - Fit for DC
    - Use your uncertainties
    - Use the right statistic
    - Accomodate your non-normal residuals

The task: finding a periodic component in a sequenc of measurements

- It may be a time series
  - Light curve: intensity vs. time
- Or a function of space
  - Temperature vs. position





Four common problems with period finding algorithms

1. Subtracting DC (instead of fitting)
2. Ignoring individual uncertainties
3. Computing the wrong statistic
   - Most commonly, using s_A^2/s_T^2 instead of s_A^2/s_E^2 as the F-statistic
4. Assuming gaussian residuals

0. Background
1. Subtracting DC (instead of fitting)
2. Ignoring individual uncertainties
3. Computing the wrong statistic
4. Assuming gaussian residuals

Use your words

- Uncertainty: the 1σ unknowable noise in your measurements
- DC: the constant a_0 in the linear model

$$y(t) = b_0 + b_1 f_1(t) + \ldots$$

- Residuals: what's left after subtracting the fit
- Errors: we don't use that word
  - It's too vague, and therefore confusing

Periodicity: The gopher in the garden

- Sometimes, you know the gopher is there, even if you can't see him
  - You can see his residue
  - But you can't tell what color his coat is
- Periodicity is the same way
  - You may be confident there is *some* periodicity, even if you can't describe it

Linear fit (aka linear regression)

- Sinusoidal fits are linear regression
  - But Lomb-Scargle uses nonlinear algorithm
- PDM *is* linear regression
- Analysis of Variance (ANOVA) is linear regression
- The "master equation" of linear regression:
  the sum-of-squares identity:

$$SST = SSA + SSE$$

total variation	modeled variation	unmodeled variation + noise

  - Absolute mathematical identity for *linear* least-squares fit

Gaussian noise special case

- For pure, gaussian noise, SSA and SSE also provide *independent* unbiased estimates of the population noise variance

$$\underbrace{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}_{SST}=\underbrace{\sum_{i=1}^{n}\left(y_{\text{mod},i}-\bar{y}\right)^2}_{SSA}+\underbrace{\sum_{i=1}^{n}\left(y_i-y_{\text{mod},i}\right)^2}_{SSE}$$

- Phase Dispersion Minimization (PDM) is a "single-treatment *p*-level ANOVA"
- Deprecated Lomb-Scargle is a 2-parameter (p = 2) linear function fit

Notation confusion

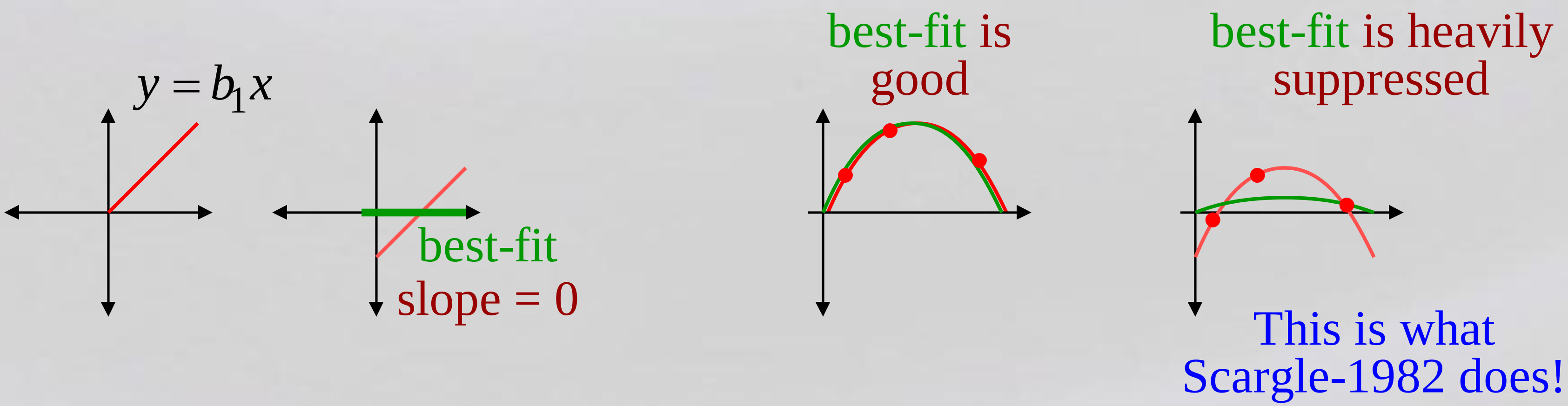
- Some references use $k \equiv$ number of parameters
- Some use $p = k + 1 \equiv$ number of parameters
  - I use this
- This affects how we write the dof
  - But not what the dof actually is

Fit for DC

0. Background
1. Fit for DC
2. Ignoring individual uncertainties
3. Computing the wrong statistic
4. Assuming gaussian residuals

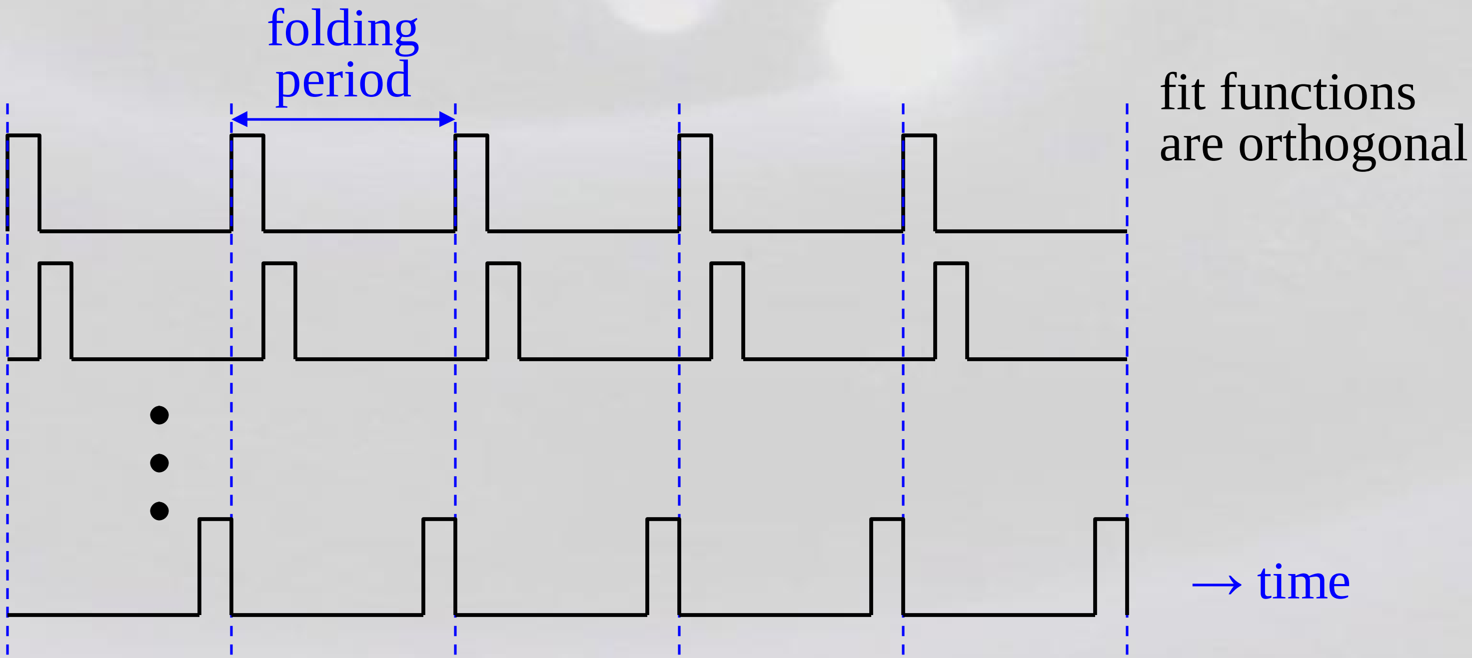
DC matters (for sinusoids)

- DC? Can't I just subtract it and forget it?
  - No
  - Simultaneous fit (DC and signal) is better than one-at-a-time with subtraction
- You would never fit a straight line without fitting for DC
  - Why would you do it with a cosine?



DC doesn't matter (for PDM)

- Not a theoretical issue with PDM
  - Round-off error usually insignificant
    - I disagree with Swarzenberg-Czerny on this
- The fit-functions subsume a DC component
  - I.e., DC is redundant with the other functions



Uncertainty Matters

0. Background
1. Fit for DC
2. Use your uncertainties
3. Computing the wrong statistic
4. Assuming gaussian residuals

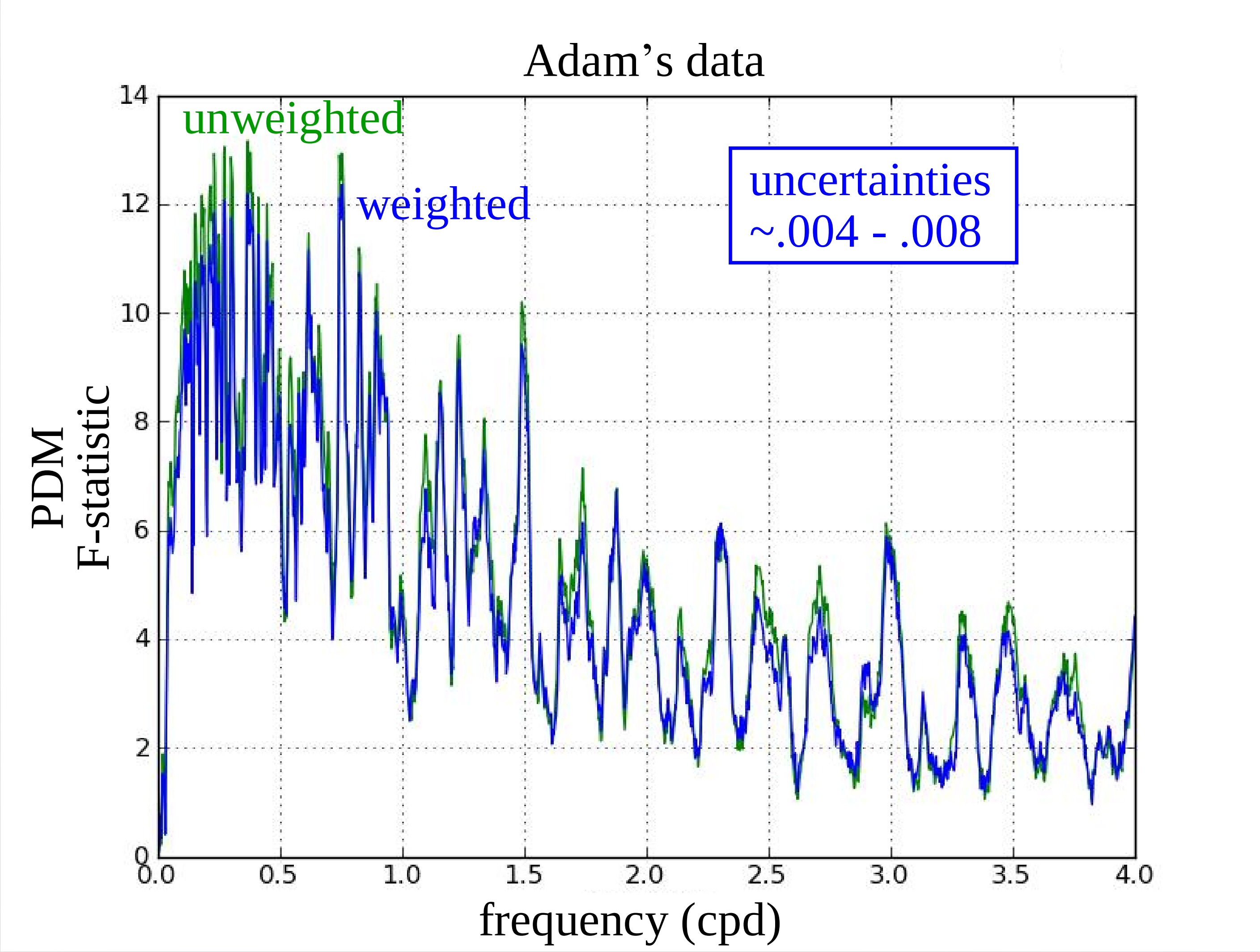
Uncertainty matters

- Just putting weights in the "obvious" places doesn't always work
  - Some things aren't obvious
  - E.g., the correct weighted sample variance is complicated:

$$s^2 \equiv \frac{\sum_{i=1}^{n} w_i \left(y_i - \overline{y} \right)^2}{V_1 - V_2 / V_1} \quad \textit{where} \quad V_1 \equiv \sum_{i=1}^{n} w_i, \quad V_2 \equiv \sum_{i=1}^{n} w_i^{\,2}$$

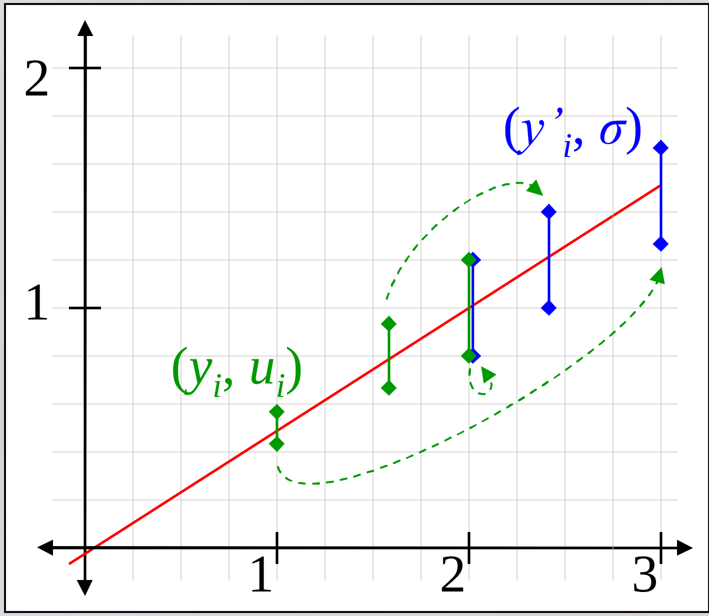
  - Doesn't work for (deprecated) Lomb-Scargle, either
    - Uses a non-linear "trick" to compute a linear fit
  - I saw two papers that used incorrect (but "obvious") formulas

Unweighted vs. Weighted



Transformation from heteroskedastic to equivalent homoskedastic measurements

- Scale *both* measurements and predictors by 1/uncertainty
- Allows using standard (homoskedastic) library routines
- Doesn't work for nonlinear algorithms
  - E.g., Lomb-Scargle doesn't work
    - τ nonlinear in y_i
    - OK, 'cuz L-S is dead, anyway



Use the Right Statistic

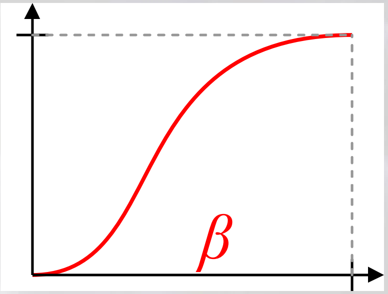
What the *F*?

0. Background
1. Fit for DC
2. Use your uncertainties
3. Use the proper F statistic
4. Assuming gaussian residuals

Correct and stable

- The PDF for the traditional Lomb-Scargle and PDM detection parameters is a beta( )
  - *Not* F !

$$\Theta_\beta \equiv \frac{SSA/dof_A}{SST/dof_T}$$

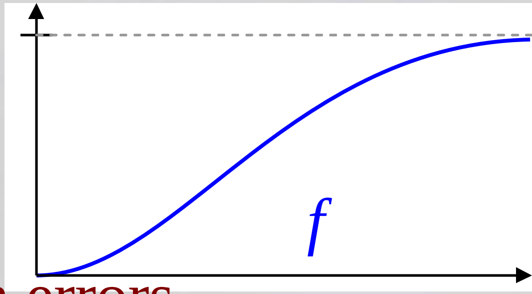


similar to

$$r^2 \equiv \frac{SSA}{SST}$$

  - But... beta( ) is numerically unstable *near the critical values*
    - Right where we need it to be accurate
- I recommend using the *correct* F statistic, which is well-behaved [Schwarzenberg-Czerny]

$$f \equiv \frac{SSA/dof_A}{SSE/dof_E}$$



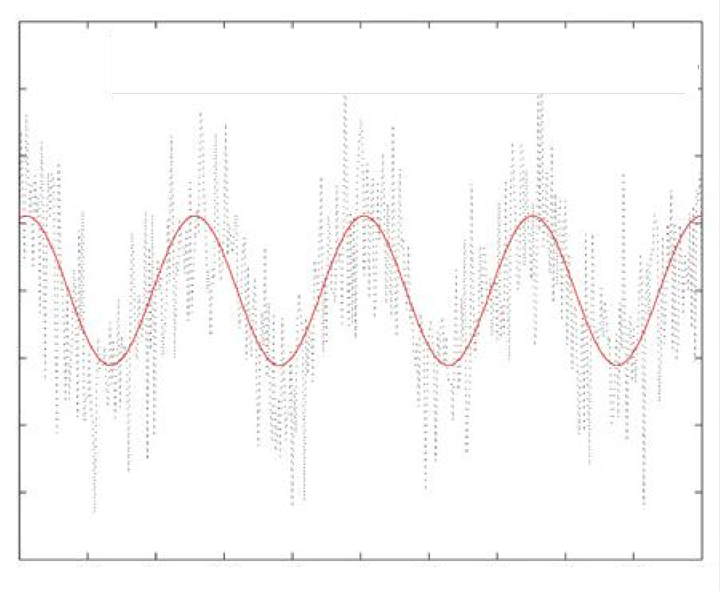
  - Pure noise $\rightarrow f \approx 1$
  - *Not* the incorrect “F” statistic in [Scargle 1982]
- Shuffle simulations hide some of these common errors
  - But it’s best to do it right, and *also* do shuffle simulations

Lomb-Scargle was a good horse while it rode ...

- But it’s got 3½ broken legs
  - It doesn’t include DC: leaves extra noise
  - It doesn’t handle uncertainties
  - Scargle recommended the wrong detection statistic

But sinusoidal periodograms live on

- The concept is perfectly valid
  - The periodogram concepts of my previous journal club are still relevant
- Zechmeister's "Modified Lomb-Scargle" is just a weighted linear fit, straight out of the book
  - E.g., Bevington
  - Crucially, it includes DC and uncertainties
  - Lomb-Scargle is dead
- I call it CSD:
  - Cosine
  - Sine
  - DC

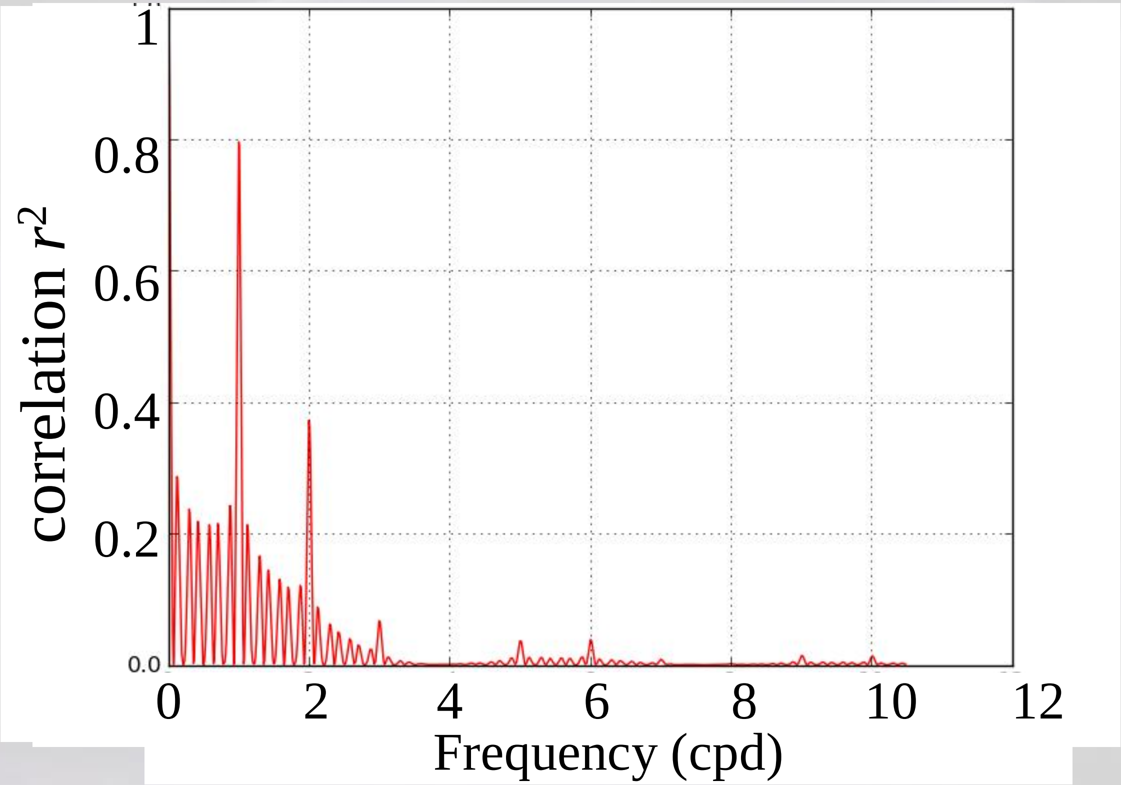


- Cosine + sine is usually not convenient
  - People want *amplitude* and *phase*, not cosine & sine
  - Especially, people want the *uncertainty* in phase
    - Can approximate *U*(phase) from *U*(cosine) and *U*(sine)

$$\mathrm{pdf}_{x,y}(x,y) \neq \mathrm{pdf}_x(x)\,\mathrm{pdf}_y(y)$$

Cycles of dependence: spectral window function

- Relevant to sinusoidal detection: CSD
  - Or DFT, and deprecated L-S
  - Only indirectly connected to PDM and Epoch-Folding



Escape From Normalcy

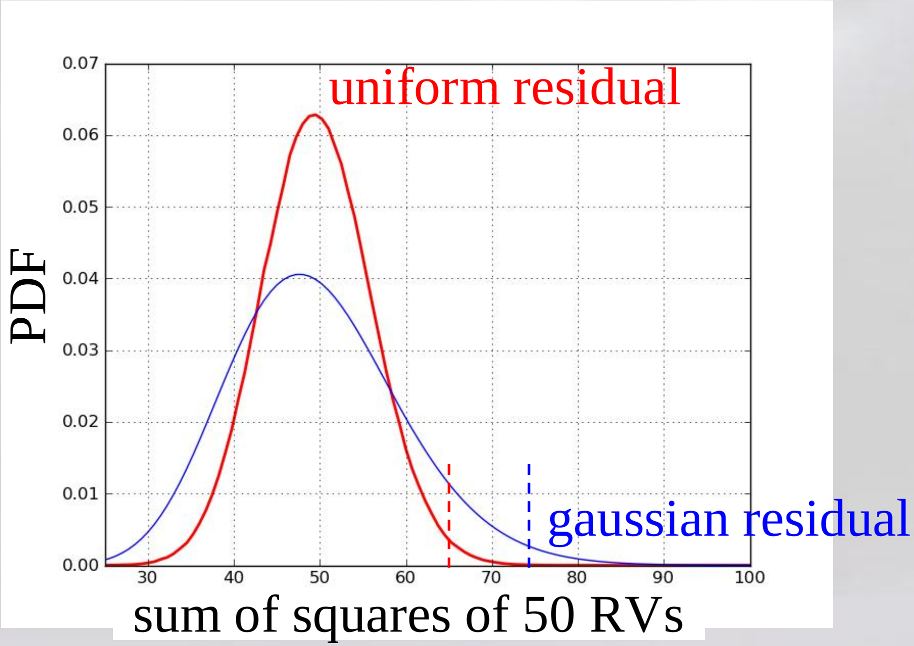
0. Background
1. Fit for DC
2. Use your uncertainties
3. Use the proper F statistic
4. Shuffle for your critical values

Your residuals aren't normal

- It's not personal
  - Nobody's are
- But it's OK
  - Your residuals are fine, just the way they are
- The critical values *must* be determined from your actual data
  - The Astronomy Shuffle

Comparison of gaussian vs. uniform residuals

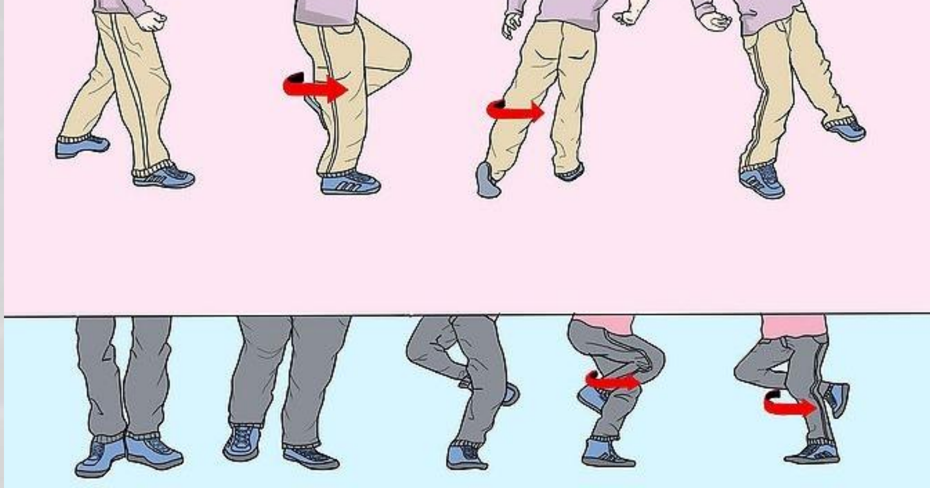
- Non-gaussian squared residuals are not χ^2
  - Using gaussian-based statistics will mask otherwise detectable signals



The Astronomy Shuffle

- The purpose of shuffling is to determine the critical values for declaring a detection from your actual data
  - You set your critical values (thresholds) according to your chosen p_{FA} (aka α)
  - The purpose is not to examine the spectral window function

The shuffle process (1)



- Randomize the order of your measurements
  - But not predictors
  - Uncertainties that are independent of measurements remain fixed
    - Uncertainties tied to measurements (e.g., photon counts) get shuffled with the data
    - Combinations should be separated into independent & dependent parts
  - Removes any periodicity from the data
    - Be careful not to use the "look for an opening" algorithm
  - Fast: $O(n)$ operations
- Repeat the following 1,000 to 10,000 times:
  - Compute a periodogram from randomized data
    - Save the max detection parameter, D_{max}, in a list
    - Different frequencies are dependent, so we can only count on one independent frequency per periodogram (slow)
- Then ...

```
 :
1.4
0.8
1.6
0.7
1.2
 :
```

The shuffle process (2)

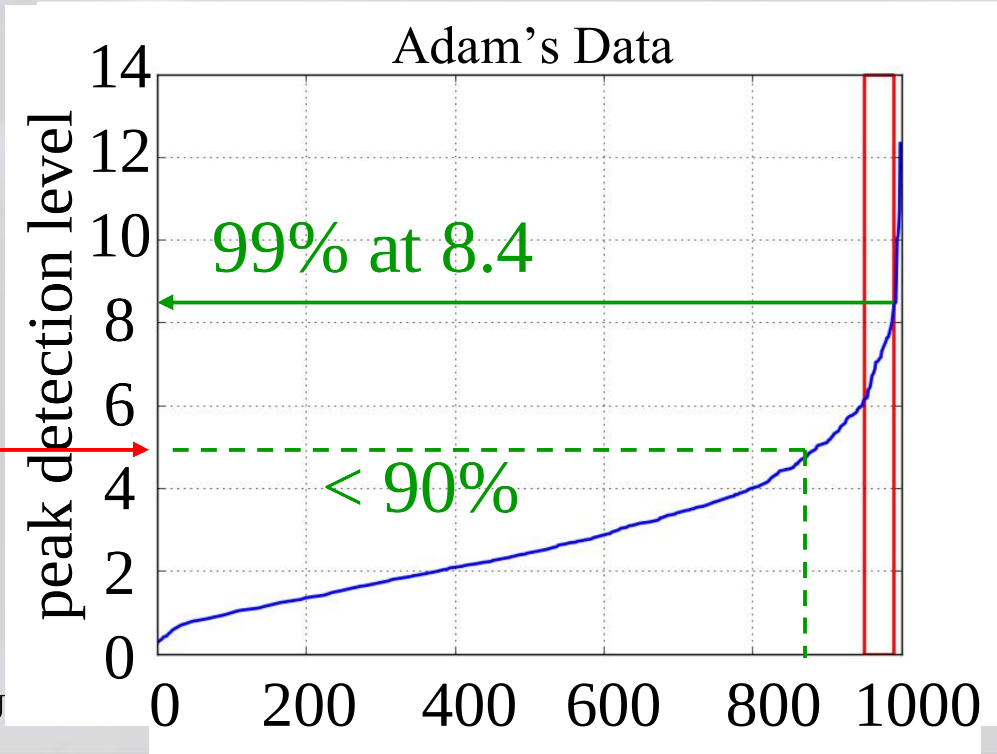
p_{FA} = 1%

99%

8.5
8.3
7.9
:
1.2
0.8
0.7
0.4

- Sort list of D_{max}
- Find your D_{critical} from p_{FA} using the sorted list
- Shuffling covers-up a lot of statistical processing errors
  - But it is best to use proper F-ish statistics, *and* shuffle simulations

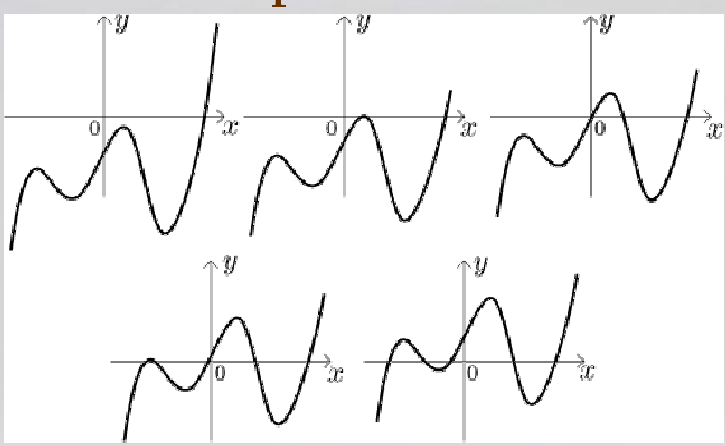
99% at 5 for gaussian noise

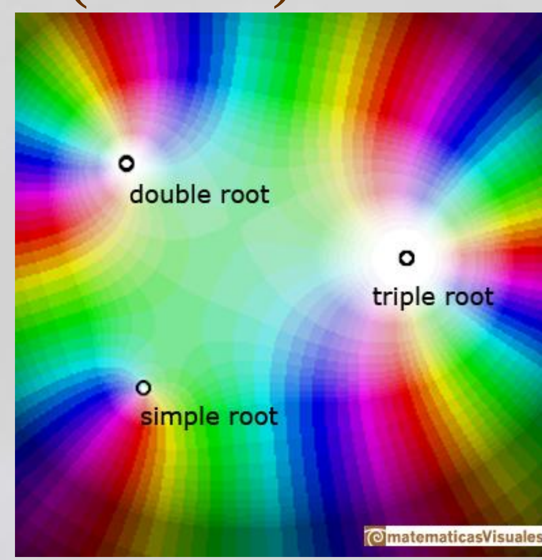


Future Directions

Future directions

- Schwarzenberg-Czerny has a paper on complex orthogonal polynomials as basis functions
  - Similar to fitting for the first few components of a Fourier Series
  - Allows for uncertainties and DC
  - You choose how many components you want (~3 or 4)
  - Small computational cost: similar to CSD





Summary

- Linear fits: Use your uncertainties
  - 3-parameter CSD fit (cosine, sine, DC) replaces Lomb-Scargle
    - Fit for DC
    - Stabilizes and speeds search for best-fit sinusoid over *freq*
    - Also provides real power, real phase
    - Lomb-Scargle is dead
  - Phase Dispersion Minimization is a non-sinusoidal period-finding algorithm
    - So is Epoch Folding
- Use the correct F-statistic: $f \equiv (SSA/dof_A)/(SSE/dof_E)$
    - Not the wrong one in [Scargle 1982] and other PDM papers
- Use shuffle simulations for your critical values
- Other tips:
  - Use simultaneous fits for multiple frequencies
    - Don't subtract one frequency at a time
  - (From last time) Know the difference between aliasing, window function correlation, and sidebands
    - These speak to the underlying physics

Thank you to ...

- Eve Armstrong for helpful comments on my practice presentation

PDM vs. Epoch Folding vs. Lomb-Scargle

- PDM and EF respond at all subharmonics
  - CSD doesn't
- PDM and EF respond to any waveform
  - CSD responds to sinusoidal components
  - Most realistic periodic signals have strong sinusoidal components
    - Exceptions include planet transits

My recommendations for period finding of smooth-signal

1. Make dense CSD periodogram
   - n_f ~= 2n, but max of 4000
   - You *have* to shuffle to determine critical values
     - The pure noise critical values are way off
2. Fit and subtract a sinusoid near the CSD peak f_1
3. From residuals, make another CSD periodogram
   - Note the peak, f_2
4. From *original* data, *simultaneously* fit two sinusoids starting from f_1 and f_2
   - *Simultaneous* regression is better than step-wise
5. From these residuals, look at the periodogram

My recommendation for finding a non-smooth signal (e.g., transit)

- Make dense PDM periodogram
  - n_f ~= 2n, but max of 4000
- Look at the folded (aka "phased") signal
  - Do any patterns catch your eye?